DOCUMENT RESUME

ED 135 813 95 TM 005 957

AUTHOR Hannan, Michael T.
TITLE Aggregation Gain Reconsidered. Technical Report No. 8.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 76
CONTRACT NIE-C-74-0123
NOTE 18p.; Paper presented at the annual meeting of the American Educational Research Association (San Francisco, April 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Error Patterns; Multiple Regression Analysis; *Research Methodology; *Statistical Bias; True Scores
IDENTIFIERS *Aggregation; Confounding Variables

ABSTRACT

Aggregation, or grouping, is a statistical procedure through which all members of a study within a specified range of scores (usually observed scores) are assigned a common or "group" score (for example, the group mean). The various social science methodology literatures agree on the costs of grouping: not only does one always lose information in grouping, in a wide variety of situations grouping introduces systematic error (bias). For most educational research applications the existing guidelines are probably appropriate. There is however, a class of situations in which grouping (of a particular type) will tend to compensate for errors in the original specification. That is, there are certain situations in which grouping produces a gain. Two special cases in which grouping is beneficial are discussed. Both cases involve estimations of effects in structural equation models. One case concerns grouping that minimizes (grouped) variation in confounding variables. The second case concerns the effects of grouping on measurement error. The benefits of the second case are less clear than in the first. The mathematical framework for both cases is presented as are areas for further investigation. (JKS)

Aggregation Gain Reconsidered*

Michael T. Hannan
Stanford University

Technical Report No. 8

April, 1976

Consortium on Methodology for Aggregating
Data in Educational Research

*This research was conducted under National Institute of Education contract #NIE-C-74-0123.

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

AGGREGATION GAIN RECONSIDERED*

Michael T. Hannan
Stanford University

Paper presented at the annual meetings of
the American Educational Research Association,
San Francisco, April 1976

*The research reported here was supported by NIE research contract C-75-0123. The paper was stimulated by comments by Francois Nielsen and Carlyle Mawon earlier work.

Aggregation Gain Reconsidered

1. Introduction

A large social science methodology literature has made clear the hazards to correct inference occasioned by using grouped data. Researchers in sociology (Robinson 1950; Goodman 1957; Blalock 1964; Hannan 1971), political science (Alker 1969; Shively 1970), economics (Theil 1954; Feige and Watts 1972), and education (Burstein 1975; Haney 1975) have all stressed the ways in which inferences from grouped observations may differ systematically from those drawn from analysis of individual (or, more generally, micro) data. There is ample evidence that the magnitude of the grouping or aggregation bias is likely to be large enough to produce very misleading findings.

Recent methodological treatments of the grouped estimation problem have focused on the ways in which the nature of the grouping process (the rule that allocates micro observations to groups) affects the divergence between grouped and ungrouped estimators (cf. Hannan and Burstein 1974; Burstein 1975). None of these treatments are general. Rather, they (following Blalock 1964) consider a variety of simple cases. These cases include random grouping and grouping that maximizes between group variation in one of the variables in a structural equations model. In each of the cases studied, grouping leads to a loss of information and consequently to a loss of efficiency in estimation. Some types of grouping processes yield estimators that contain an aggregation bias, while others do not. In none of these cases is there any gain from grouping.

In an early and important paper, Grunfeld and Griliches (1960) proposed that grouping may in some cases lead to a gain. They considered the effect of grouping on estimators of R^2 from micro models that are improperly specified. They pointed out the possibility that the grouping bias might offset the

specification bias in such a way that the R^2 calculated from grouped data might be closer to the true R^2 than that calculated from an incorrect micro model. Hannan and Burstein (1974) studied this issue with reference to estimators of structural parameters (e.g. path regressions). For the range of grouping cases they considered they found no evidence of any aggregation gain. They did identify cases where grouping magnifies specification bias in the micro model and others where there is no magnification, but none where there was a reduction. While their argument appears correct as it stands, it gives a misleading impression that grouping will never yield gains in terms of bias.

My purpose is to reopen the issue of aggregation gain. I will show that the simpler framework used by Hannan et al. (1975) to relate grouping effects to specification bias makes clear that aggregation gain is possible. Then I will explore two interesting cases.

2. Framework

At a minimum we must consider a (true) model and two alternative estimators: ungrouped and grouped. We want to compare the properties of the estimators (here only their means) under various types of grouping processes and various types of model misspecification. For example, consider the following model:

$$y = \beta_1 x_1 + \beta_2 x_2 + u$$

$$\text{or} \quad \underline{y} = \underline{X}\underline{\beta} + \underline{u} \tag{1}$$

where X_1 and X_2 are stochastic regressors[1] and

$$\text{plim}\left(\frac{1}{N}\underline{u}'\underline{u}\right) = \sigma_u^2$$

$$\text{plim}\left(\frac{1}{N}\underline{X}'\underline{X}\right) = \Sigma$$

$$\text{plim}\left(\frac{1}{N}\underline{X}'\underline{u}\right) = 0$$

(where plim denotes the probability limit, cf. Johnston 1972: 268-281).

We consider the usual ungrouped ordinary least squares estimator:

$$\underline{\hat{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y} \qquad (2)$$

and a grouped estimator

$$\underline{\bar{\beta}} = (\underline{\bar{X}}'\underline{\bar{X}})^{-1}\underline{\bar{X}}'\underline{\bar{y}} \qquad (3)$$

where the bars over vectors and matrices indicate that they contain grouped observations. Each type of grouping process determines a grouped estimator.

We might define asymptotic grouping bias as

$$\text{plim } (\underline{\bar{\beta}} - \underline{\hat{\beta}}).$$

But, this is a meaningful criterion only when the ungrouped estimator is asymptotically unbiased. More generally, we must consider the possibility that $\underline{\hat{\beta}}$ is biased. To be concrete, we treat the estimator of β_1 that ignores the presence of x_2 in the model

$$\tilde{\beta}_1 = \frac{\Sigma x_1 y}{\Sigma x_1^2} . \qquad (4)$$

As long as plim $(\frac{1}{N}\underline{x}'\underline{x}) \neq 0$, $\tilde{\beta}_1$ is an inconsistent estimator of β, (Theil 1957). The specification bias of $\tilde{\beta}$, is defined as

$$\text{plim } (\tilde{\beta}_1 - \beta_1) = \beta_2 \frac{\sigma_{x_1 x_2}}{\sigma_{x_1}^2}$$

A grouped analogue to (4) is

$$\bar{\tilde{\beta}}_1 = \frac{\Sigma \bar{x}_1 \bar{y}}{\Sigma \bar{x}_1^2} \qquad (5)$$

with

$$\text{plim } (\bar{\tilde{\beta}}_1 - \beta_1) = \beta_2 \frac{\sigma_{\bar{x}_1 \bar{x}_2}}{\sigma_{\bar{x}_1}^2}$$

where $\sigma_{\bar{x}_1 \bar{x}_2}$ and $\sigma^2_{\bar{x}_1}$ denote population covariances and variances under the given grouping rule.

Finally, we want to compare the grouped estimator (5) with the ungrouped estimator (4). As we have constructed the example, both estimators are inconsistent. The question of aggregation gain concerns the possibility that the asymptotic bias may be smaller in the grouped estimator. A natural definition of aggregation gain is (cf. Grunfeld and Griliches; Hannan and Burstein 1974):

$$|\text{plim } (\bar{\bar{\beta}} - \beta)| < |\text{plim}(\tilde{\beta} - \beta)| \qquad (6)$$

To evaluate expressions like (6) we must take explicit account of the nature of the grouping process. In the cases we wish to consider the grouping rule (more formally, the grouping matrix in Prais and Aitchison's (1974) terminology) is stochastic. That is, the rule that places individuals in groups utilizes the outcome of some stochastic process (e.g. places individuals in groups on the basis of their value on one or more of the variables in the model). This complication makes it very difficult to obtain exact expressions for (6). We continue to use large sample results (probability limits) and particularly Monte Carlo results to evaluate (6) for the cases of interest. The relevant Monte Carlo results are presented more fully in Hannan et al. (1975) and Hannan and Young (1976). The most important finding is that the simulation results for modest sized samples (N=500 grouped by 10's) agree closely with the large sample analytic results.

3. Aggregation Gain: Omitted Variables

The first case we investigate is that set out in the previous section: specification bias due to the omission of a causal variable related to the included causal variable. The ungrouped estimator from (4) is biased and inconsistent.

$$\text{plim } \hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sigma_{x_1 x_2}}{\sigma_{x_1}^2}$$

as is the grouped estimator:

$$\text{plim } \bar{\beta}_1 = \beta_1 + \beta_2 \frac{\sigma_{\bar{x}_1 \bar{x}_2}}{\sigma^2_{\bar{x}_1}}$$

So aggregation gain requires that

$$\left| \frac{\sigma_{\bar{x}_1 \bar{x}_2}}{\sigma^2_{\bar{x}_1}} \right| < \left| \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}} \right| \qquad (7)$$

As we reported earlier (Hannan et al. 1975) none of the commonly studied cases meets this criterion. However, from (7) it is clear that certain types of grouping rules will yield an aggregation gain. For example, any grouping process that eliminates the covariance of x_1 and x_2 in the grouped data will yield such a gain. What would such a grouping process look like? One simple case is a grouping that eliminates between group variation in x_2. Then the grouped estimator is unbiased while the ungrouped estimator is not.

This case is not completely arcificial. Consider the following concrete example. Davis (1966) proposed that student aspirations for additional educational attainment respond to a "frog-pond" effect. The higher the level of performance of one's peers holding constant one's own performance level, the lower are one's aspirations. Suppose that the effect operates more precisely as follows. Let aspirations (y) depend linearly on performance level (x_1) and rank in class (x_2). Analyses that ignore x_2 will give biased estimates of the performance effect when individual data are used but not when class averages are employed.

There is a general class of situations of which this one is an example. Rank in class is a variable defined relative to some bounded system

(a "relational variable" in Lazarsfeld and Menzel's (1965) terminology). Whenever such variables are omitted from a model and the grouping corresponds with the system boundaries (so that there is no between-group variance in the relational variable), grouping will produce a gain.

4. Aggregation Gain: Measurement Error

The Hannan-Burstein and Grunfeld-Griliches analyses presume perfect measurement of causal variables. In this section we address the possibility of aggregation gain in simple models in which the causal variables are measured with random error. In particular, we use the following model

$$y = \beta x + u$$

$$x' = x + \varepsilon$$

$$\text{plim } (\tfrac{1}{N} \Sigma xu) = \text{plim } (\tfrac{1}{N} \Sigma x\varepsilon) = \text{plim } (\tfrac{1}{N} \Sigma u\varepsilon) = 0$$

The ungrouped estimator of interest is

$$\hat{\beta} = \frac{\Sigma x'y}{\Sigma x'^2}$$

and (cf. Johnston 1972: 282)

$$\text{plim } \hat{\beta} = \frac{\beta}{1 + \sigma_\varepsilon^2/\sigma_x^2}$$

That is, the ungrouped estimator contains an asymptotic specification bias that depends on the ratio of measurement error variance to true score variance.

Next, we consider two grouping processes and the resulting grouped estimators: (1) grouping that maximizes grouped true score variance; and (2) grouping that maximizes observed score variance.

A. Grouping that maximizes grouped true score variance

From our previous work, we know that grouping that maximizes grouped variation

in x is random with respect to ε. In fact, under these conditions we found that

$$\sigma^2_{\bar{x}} \cong \sigma^2_x \quad \text{and} \quad \sigma^2_{\bar{\varepsilon}} = \sigma^2_\varepsilon/n$$

where n is the size of each group (assuming equal-sized groups). Using these results as an approximation we have

$$\text{plim } \bar{\beta} = \text{plim} \left[\frac{\Sigma \bar{x}'\bar{\varepsilon}}{\Sigma \bar{x}'^2}\right] \cong \frac{\beta}{1 + \sigma^2_\varepsilon/n\sigma^2_x}$$

Clearly with these approximations, there is an aggregation gain. For example, in our simulation with the reliability of x' = .7 and groups of size 10

$$\text{plim } \hat{\beta} \cong .54\beta \; ; \; \text{plim } \bar{\beta} \cong .92\beta$$

when the reliability is .3

$$\text{plim } \hat{\beta} \cong .18\beta \; ; \; \text{plim } \bar{\beta} \cong .69\beta$$

As we would expect, the lower the reliability the greater the aggregation gain.

These sorts of considerations prompted Wald (1940) and Bartlett (1949) to propose certain grouped estimators as improvements over the usual ungrouped estimators. They failed to realize, however, that the estimators they proposed are consistent only when the observations are grouped by true scores (Neyman and Scott, 1954). I have not yet been able to identify a realistic situation in educational research in which observations are grouped by true scores.[2] Therefore, it is important to investigate the consequences of grouping by observed scores (x').

B. Grouping by fallibly measured scores

In realistic situations, observations are grouped by observed scores. Here we consider the analogue to the case just discussed, namely, grouping that maximizes between groups variation in x'. An additional complication arises

in this case since according to the model x' is endogenous (causally dependent). As Blalock (1964) and others have demonstrated, grouping by values of endogenous variables tends to produce a (positive) correlation between regressors and disturbances even when they are independent in the ungrouped data (cf. Hannan and Young, 1976 for Monte Carlo evidence on this). As a consequence we cannot presume in this case that grouped true scores ($\bar{x}$) and grouped measurement errors ($\bar{\epsilon}$) will be uncorrelated (even asymptotically). That is, the grouped estimator has the following asymptotic bias:

$$\text{plim } \bar{\beta} = \text{plim}\left(\frac{\Sigma \bar{x}'\bar{y}}{\Sigma \bar{x}'^2}\right) = \frac{\beta}{1 + \sigma^2_{\bar{\epsilon}} \Big/ \left(\sigma^2_{\bar{x}} + \sigma_{\bar{x}\bar{\epsilon}}\right)}$$

The comparison of grouped and ungrouped estimators is more complicated than in the previous case. Asymptotic aggregation gain requires that

$$\frac{\sigma^2_{\epsilon}}{\sigma^2_{x}} > \frac{\sigma^2_{\bar{\epsilon}}}{\sigma^2_{\bar{x}} + \sigma_{\bar{x}\bar{\epsilon}}}$$

or

$$\sigma_{\bar{x}\bar{\epsilon}} > \left(\frac{\sigma^2_{\bar{\epsilon}}}{\sigma^2_{\epsilon}}\right)\sigma^2_{x} - \sigma^2_{\bar{x}}$$

As far as I have been able to determine, this condition is not inconsistent with the model specification and grouping process. Our simulation (conducted only on three variable models) does not yield the quantities necessary to evaluate the possibility of aggregation gain in small samples for this type of grouping.

Blalock et al. (1970) report a Monte Carlo study that is relevant here. They compared the behavior of the Wald and Bartlett estimators (with data

grouped by observed scores) with the ungrouped ordinary least squares estimator. These grouped estimators are different from the estimator just discussed but are roughly analogous. At any rate, Blalock et al. found no gain over the ungrouped estimator. In all cases simulated, the behavior of the Wald and Bartlett estimators was quite similar on the average to the ungrouped estimator. We will shortly revise our simulation to conduct a systematic study of the question of aggregation gain under these conditions.

Finally, we note an interesting attempt by Aigner and Goldfeld (1974) to explicate the original Grunfeld-Griliches argument. As in most of the economics literature, the problem is viewed from the perspective that the micro relations differ from individual to individual. Consequently, if there are N individuals, there are N structural equations to be estimated. We have been considering the simpler case where all micro units are assumed to behave according to the same structural relationship. Aigner and Goldfeld do treat this problem as a special case. In so doing, they pose a clear example of aggregation gain. The micro model has the form (a time series):

$$y_1 = \beta x_1 + u_1$$

$$y_2 = \beta x_2 + u_2$$

i.e., $$y_i = \beta x_i + u_i$$

and x_i are unobserved, and are replaced by indicators measured with random error. In this extreme case, the random measurement errors are equal and opposite in sign:

$$y_1' = y_1 + \epsilon$$

$$y_2' = y_2 - \epsilon.$$

The grouped model is

$$(y_1 + y_2) = \beta(y_1' + y_2') + W.$$

Note that the random errors cancel in the grouped data. Consequently, the grouped estimator will be consistent while ungrouped ordinary least squares estimators will not.

The example is obviously artificial. Nontheless, it does give a clear indication that under at least some conditions grouping may lead to an aggregation gain in models that suffer from errors in variables even when one cannot group by true scores.

5. Conclusions

The various social science methodology literatures agree on the costs of grouping. One always loses information in grouping. Moreover, in a wide variety of situations grouping introduces systematic error. For most educational research applications the existing guidelines are probably appropriate. There is, however, a class of situations in which grouping (of a particular type) will tend to compensate for errors in the original specification. That is, there are certain situations in which grouping produces a gain.

We have made the case for aggregation gain by examining two special cases. The first involves grouping that minimizes (grouped) variation in confounding variables. Obviously if grouping can eliminate such variation it may improve inference. We have shown that when the confounding variables are relational (defined relative to the group), grouping may yield a gain. The second case concerns the effect of grouping on measurement error. As has been widely recognized, grouping by true scores will yield a gain relative to estimators that employ ungrouped fallibly measured variables. The more realistic case

in which observations are grouped by observed scores is more complicated. However, it appears that aggregation gain is possible in this case as well. At least we cannot on the basis of existing results rule out this possibility.

In summary, we argue against overgeneralizing the results on the costs of aggregation. Whether or not grouping yields costs or gains cannot be determined without knowledge of the process that groups observations and the nature of the substantive problem and research design. No methodological guideline substitutes for careful scrutiny of each application.

Footnotes

[1] Much of the literature on grouped estimation considers nonstochastic regressors. Since, as we point out below, the grouping matrices we consider are stochastic, the grouped regressors are stochastic. As a consequence, there is nothing to be gained by preserving the assumption that the ungrouped regressors are fixed. The presence of stochastic regressors forces us to use weaker results than for the fixed case. In particular, we examine probability limits of estimators.

[2] I presume that the true scores are unknown. Otherwise a rational investigation would not use the ungrouped estimator considered here.

References

Aigner, D.J. and S.M. Goldfeld
1974 "Estimation and prediction from aggregate data when aggregates are measured more accurately than their components," Econometrics 42: 113-134.

Alker, Hayward R., Jr.
1969 "A typology of ecological fallacies," in Quantitative Ecological Analysis in the Social Sciences. Edited by M. Dogan and S. Rokkan, Cambridge, Mass.: MIT Press, pp. 69-86.

Bartlett, M.S.
1949 "Fitting a straight line when both variables are subject to error," Biometrics 5: 207-212.

Blalock, Hubert M., Jr.
1964 Causal Inferences in Nonexperimental Research. Chapel Hill, N.C.: University of North Carolina Press.

Blalock, Hubert M., Jr., C. Wells, and L. Carter
1970 "Statistical estimation with random measurement error," in E. Borgotta and G. Bohrenstedt (eds.) Sociological Methodology 1970. San Francisco: Jossey-Bass.

Burstein, Leigh
1975 The Use of Data from Groups for Inferences about Individuals in Educational Research. Stanford University unpublished Ph.D. dissertation.

Davis, James
1966 "The campus as a frog-pond," American Journal of Sociology 72: 17-31.

Feige, Edgar L. and Harold W. Watts
1972 "An investigation of the consequences of partial aggregation of micro-economic data," Econometrics 40(March): 343-360.

Goodman, Leo
1959 "Some alternative to ecological correlation," American Journal of Sociology 64(May): 610-625.

Grunfeld, Yehuda, and Griliches, Zvi
1960 "Is aggregation necessarily bad?" Review of Economics and Statistics 42 (February): 1-13.

References

Haney, Walt
1974 Unit of Analysis Issues in the Evaluation of Project Follow-Through: Or There Must be Heresy in This Some Place. Huron Institute.

Hannan, Michael T.
1971 Aggregation and Disaggregation in Sociology. Lexington, Mass.: Heath-Lexington.

Hannan, Michael T. and Leigh Burstein
1974 "Estimation from grouped observations," American Journal of Sociology 39(June): 374-392.

Hannan, Michael T., A. Young, and F. Nielsen
1975 "Specification bias analysis of the effects of grouping of observations in multiple regression models." Paper presented at the annual meetings of American Educational Research Association.

Hannan, Michael T., J. Freeman, and J. Meyer
1976 "Specification of models of organizational effectiveness: comment on Bidwell and Kasarda," American Sociological Review 41(136-143).

Hannan, Michael T. and A. Young
1976 "Small sample results on estimation from grouped observations," unpublished manuscript.

Hanushek, Eric A., John E. Jackson, and John F. Kain
1974 "Model specification, use of aggregate data, and the ecological correlation fallacy," Political Methodology 1(August): 89-107.

Johnston, J.
1972 Econometric Methods. 2nd edition, New York: McGraw-Hill.

Lazarfeld, Paul and Herbert Menzel
1965 "On the relations between individual and collective properties," in Complex Organizations. Edited by Amitai Etzioni. New York: Rinehart and Winston, pp. 422-440.

Neyman, J. and E.L. Scott
1954 "On certain methods of estimating the linear structural relation," Annals of Mathematical Statistics 22: 352-361; and Correction: Vol. 23: 115.

References

Prais, S.J. and J. Aitchison
1954 "The grouping of observations in regression analysis," Review of the International Statistical Institute 22: 1-22.

Robinson, William S.
1950 "Ecological correlations and the behavior of individuals," American Sociological Review 15(June): 351-357.

Shively, W. Phillip
1969 "'Ecological' inference: the use of aggregate data to study individuals," American Political Science Review 63(December): 1183-1196.

Theil, Henri
1954 Linear Aggregation in Economic Relations. Amsterdam: North Holland Publishing Company.

1957 "Specification errors and the estimation of economic relationships," Review of International Statistical Institute 25: 41-51.

Wald, A.
1940 "The fitting of straight lines if both variables are subject to error," Annals of Mathematical Statistics 11: 284-300.